Support Vector Machines & Kernels
Lecture 5

David Sontag
New York University

Slides adapted from Luke Zettlemoyer and Carlos Guestrin

Support Vector Machines

QP form:

$$\underset{\mathbf{w},\xi_i \geq 0}{\text{argmin}} \ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i$$

$$\text{s.t.} \quad \forall i, \ y_i\langle \mathbf{w}, \mathbf{x}_i\rangle \geq 1 - \xi_i$$

More "natural" form:

Equivalent if $C = \frac{1}{m\lambda}$

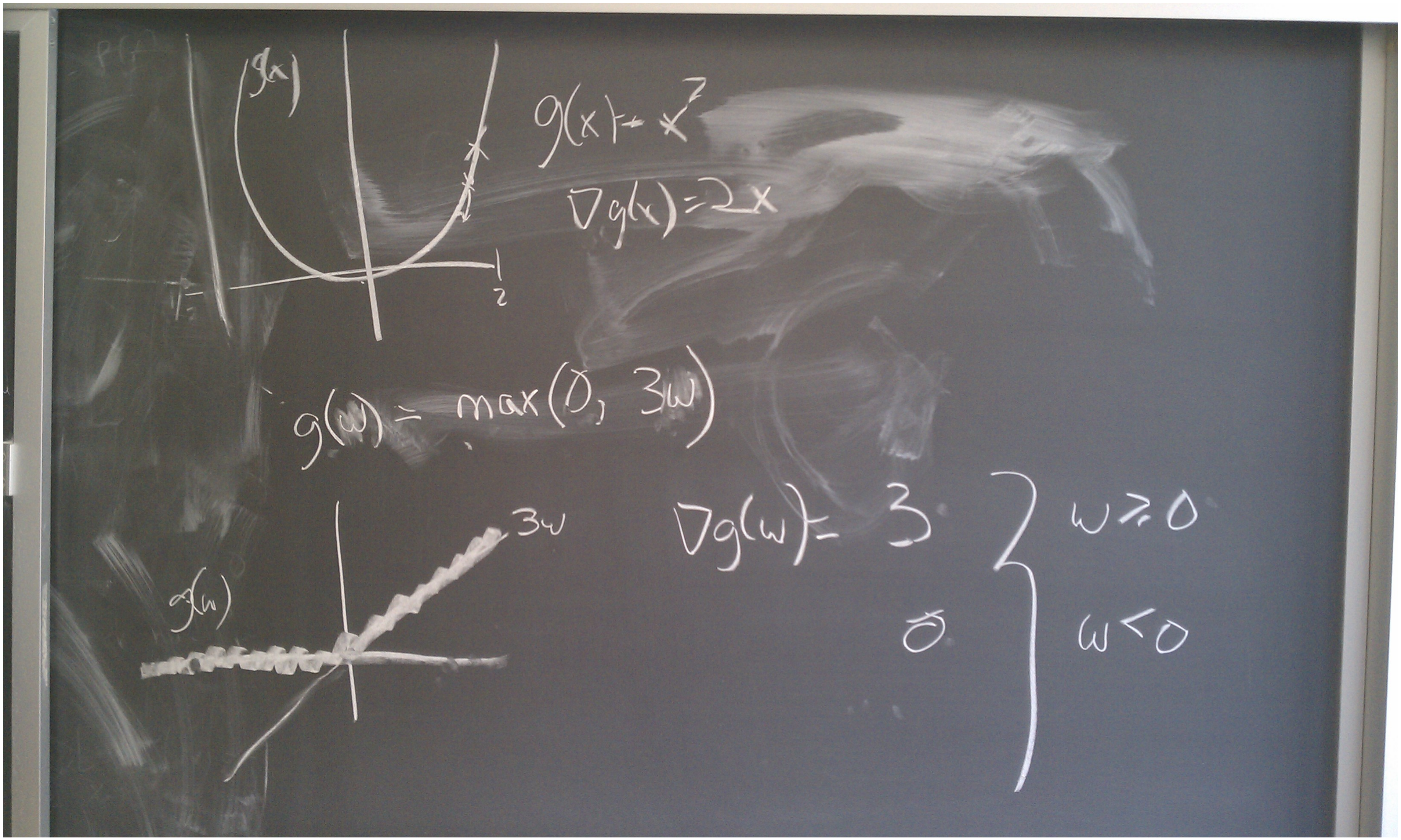
$$\underset{\mathbf{w}}{\text{argmin}} \ f(\mathbf{w}) \qquad \text{where:}$$

$$f(\mathbf{w}) \stackrel{\text{def}}{=} \frac{\lambda}{2}\|\mathbf{w}\|^2 + \frac{1}{m}\sum_{i=1}^{m}\max\{0, 1 - y_i\langle \mathbf{w}, \mathbf{x}_i\rangle\}$$

Regularization term

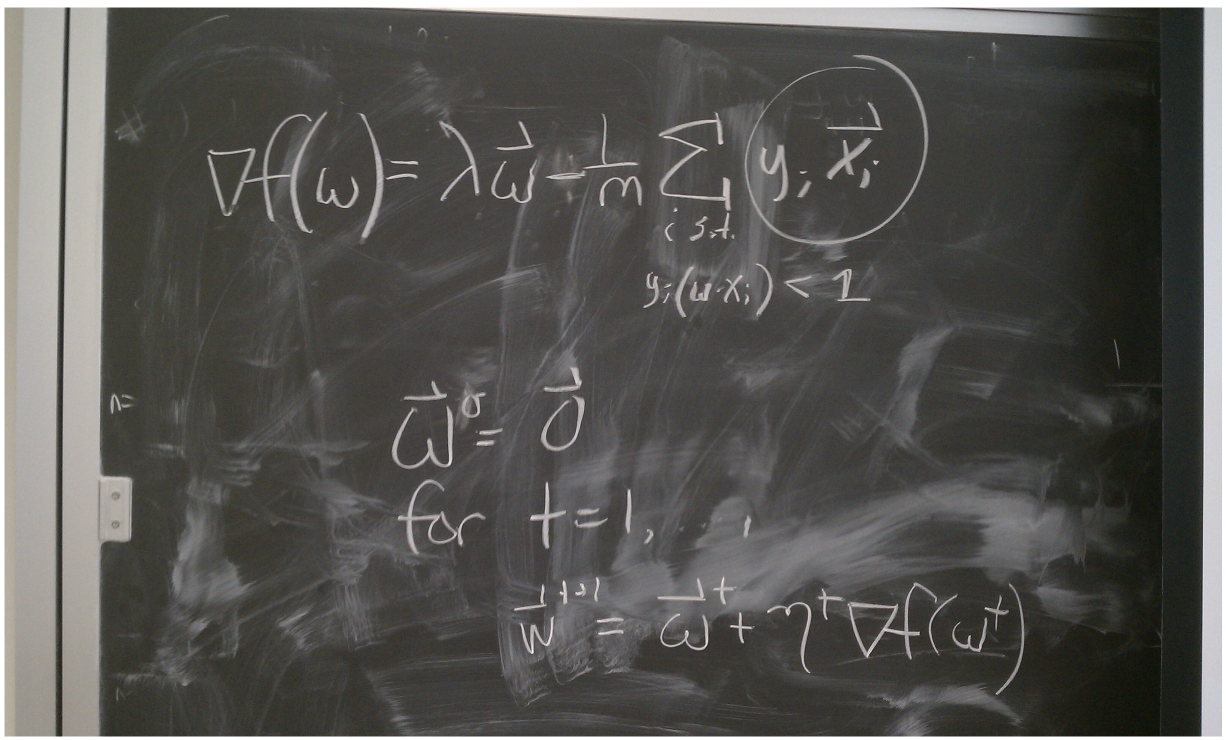
Empirical loss

Subgradient method



Subgradient method

$$f(\mathbf{w}) \stackrel{\text{def}}{=} \frac{\lambda}{2}\|\mathbf{w}\|^2 + \frac{1}{m}\sum_{i=1}^{m} \max\{0, 1 - y_i\langle \mathbf{w}, \mathbf{x}_i\rangle\}$$



Step size:

$$\eta_t = \frac{1}{t\lambda}$$

Stochastic subgradient

INPUT: training set $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$,
Regularization parameter λ,
Number of iterations T

INITIALIZE: Choose $\mathbf{w}_1$ s.t. $\|\mathbf{w}_1\| \leq 1/\sqrt{\lambda}$

FOR $t = 1, 2, \ldots, T$

Subgradient:

- Choose $A_t \subseteq S$
- $A_t^+ = \{(\mathbf{x}, y) \in A_t : y\langle \mathbf{w}_t, \mathbf{x}\rangle < 1\}$
- $\nabla_t = \lambda \mathbf{w}_t - \frac{1}{|A_t|} \sum_{(\mathbf{x},y) \in A_t^+} y\,\mathbf{x}$
- $\eta_t = \frac{1}{t\,\lambda}$
- $\mathbf{w}'_t = \mathbf{w}_t - \eta_t \nabla_t$

OUTPUT: $\mathbf{w}_{T+1}$

PEGASOS

A_t = S

Subgradient method

|A_t| = 1

Stochastic gradient

... t S = ...

... ation ...

Number of iterations T

INITIALIZE: Choose $\mathbf{w}_1$ s.t. $\|\mathbf{w}_1\| \leq 1/\sqrt{\lambda}$

FOR $t = 1, 2, \ldots, T$

Subgradient:

- Choose $A_t \subseteq S$
- $A_t^+ = \{(\mathbf{x}, y) \in A_t : y\langle \mathbf{w}_t, \mathbf{x} \rangle < 1\}$
- $\nabla_t = \lambda \mathbf{w}_t - \frac{1}{|A_t|} \sum_{(\mathbf{x},y) \in A_t^+} y\,\mathbf{x}$
- $\eta_t = \frac{1}{t\,\lambda}$
- $\mathbf{w}_t' = \mathbf{w}_t - \eta_t \nabla_t$

Projection $\Leftarrow$ $\mathbf{w}_{t+1} = \min\left\{1, \frac{1/\sqrt{\lambda}}{\|\mathbf{w}_t'\|}\right\} \mathbf{w}_t'$

OUTPUT: $\mathbf{w}_{T+1}$

Run-Time of Pegasos

- Choosing $|A_t|=1$

  ➔ Run-time required for Pegasos to find ε accurate solution w.p. $\geq$ 1-δ

$$\tilde{O}\left(\frac{n}{\delta\,\lambda\,\epsilon}\right)$$

n = # of features

- Run-time does not depend on #examples
- Depends on "difficulty" of problem (λ and ε)

Experiments

- **3 datasets** (provided by Joachims)
  - Reuters CCAT (800K examples, 47k features)
  - Physics ArXiv (62k examples, 100k features)
  - Covertype (581k examples, 54 features)

Training Time (in seconds):

	Pegasos	SVM-Perf	SVM-Light
Reuters	**2**	77	20,075
Covertype	**6**	85	25,514
Astro-Physics	**2**	5	80

What’s Next!

- Learn one of the most interesting and exciting recent advancements in machine learning
  - The “kernel trick”
  - High dimensional feature spaces at no extra cost
- But first, a detour
  - Constrained optimization!

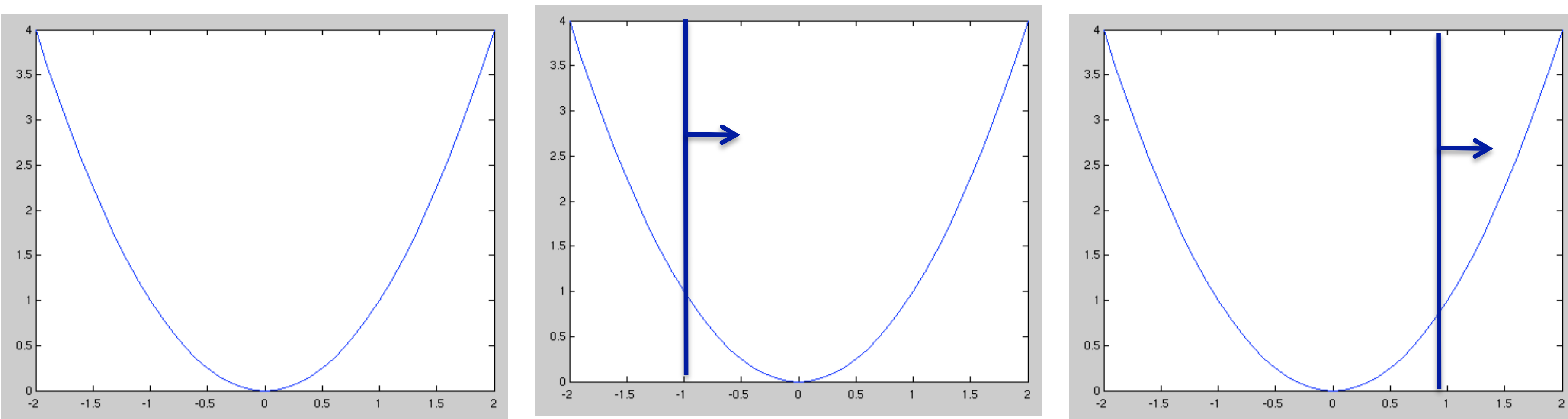
Constrained optimization

$$\min_x \; x^2 \quad \text{s.t.} \quad x \geq b$$

No Constraint

x ≥ -1

x ≥ 1



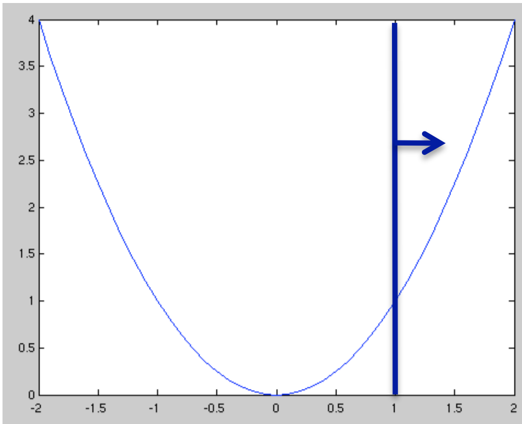
x*=0

x*=0

x*=1

How do we solve with constraints?
→ Lagrange Multipliers!!!

Lagrange multipliers – Dual variables



$$\min_x \; x^2$$
$$\text{s.t.} \quad x \geq b$$

Add Lagrange multiplier

Rewrite Constraint

Introduce Lagrangian (objective):

$$L(x, \alpha) = x^2 - \alpha(x - b)$$

We will solve:

$$\min_x \max_\alpha \; L(x, \alpha)$$
$$\text{s.t.} \quad \alpha \geq 0$$

Add new constraint

Why is this equivalent?

- min is fighting max!

x<b → (x-b)<0 → $\max_\alpha -\alpha(x-b) = \infty$

- min won't let this happen!

x>b, $\alpha \geq 0$ → (x-b)>0 → $\max_\alpha -\alpha(x-b) = 0$, $\alpha^* = 0$

- min is cool with 0, and $L(x, \alpha) = x^2$ (original objective)

x=b → α can be anything, and $L(x, \alpha) = x^2$ (original objective)

The *min* on the outside forces *max* to behave, so constraints will be satisfied.

Dual SVM derivation (1) – the linearly separable case (hard margin SVM)

Original optimization problem:

$$\text{minimize}_{\mathbf{w},b} \quad \frac{1}{2}\mathbf{w}.\mathbf{w}$$
$$\left(\mathbf{w}.\mathbf{x}_j + b\right) y_j \geq 1, \ \forall j$$

Rewrite constraints

One Lagrange multiplier per example

Lagrangian:

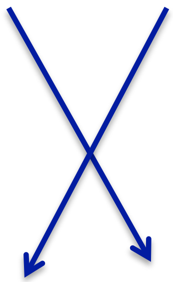
$$L(\mathbf{w}, \alpha) = \frac{1}{2}\mathbf{w}.\mathbf{w} - \sum_j \alpha_j \left[\left(\mathbf{w}.\mathbf{x}_j + b\right) y_j - 1\right]$$
$$\alpha_j \geq 0, \ \forall j$$

Our goal now is to solve: $\min_{\vec{w},b} \ \max_{\vec{\alpha}\geq 0} \ L(\vec{w}, \vec{\alpha})$

Dual SVM derivation (2) – the linearly separable case (hard margin SVM)

(Primal)

$$\min_{\vec{w},b} \max_{\vec{\alpha}\geq 0} \frac{1}{2}||\vec{w}||^2 - \sum_j \alpha_j \left[(\vec{w}\cdot\vec{x}_j + b)\, y_j - 1 \right]$$



Swap min and max

(Dual)

$$\max_{\vec{\alpha}\geq 0} \min_{\vec{w},b} \frac{1}{2}||\vec{w}||^2 - \sum_j \alpha_j \left[(\vec{w}\cdot\vec{x}_j + b)\, y_j - 1 \right]$$

Slater's condition from convex optimization guarantees that these two optimization problems are equivalent!

Dual SVM derivation (3) – the linearly separable case (hard margin SVM)

(Dual) $$\max_{\vec{\alpha}\geq 0} \min_{\vec{w},b} \frac{1}{2}||\vec{w}||^2 - \sum_j \alpha_j \left[(\vec{w}\cdot\vec{x}_j + b)\, y_j - 1 \right]$$

Can solve for optimal **w**, b as function of α:

$$\frac{\partial L}{\partial w} = w - \sum_j \alpha_j y_j x_j \quad \rightarrow \quad \mathbf{w} = \sum_j \alpha_j y_j \mathbf{x}_j$$

$$\frac{\partial L}{\partial b} = -\sum_j \alpha_j y_j \quad \rightarrow \quad \sum_j \alpha_j y_j = 0$$

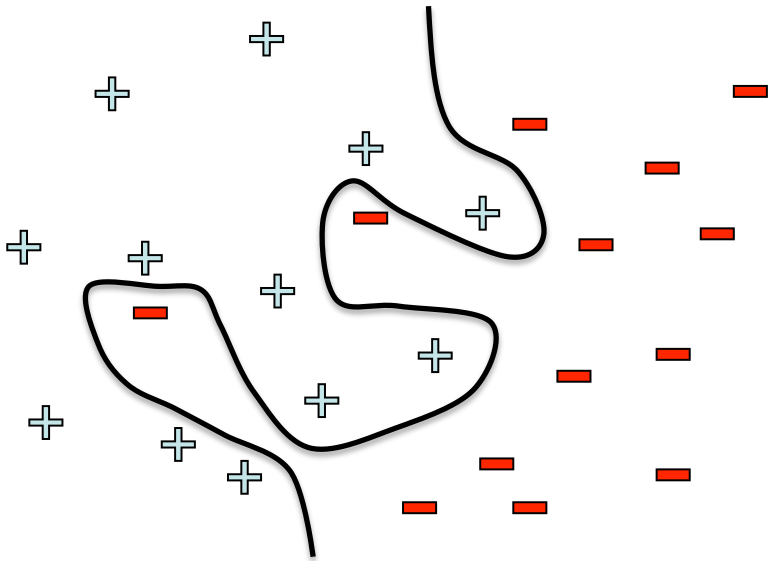
Substituting these values back in (and simplifying), we obtain:

(Dual) $$\max_{\vec{\alpha}\geq 0,\, \sum_j \alpha_j y_j = 0} \sum_j \alpha_j - \frac{1}{2}\sum_{i,j} \underbrace{y_i y_j \alpha_i \alpha_j}_{\text{scalars}} \underbrace{(\vec{x}_i \cdot \vec{x}_j)}_{\text{dot product}}$$

Sums over all training examples   scalars   dot product

Reminder: What if the data is not linearly separable?

Use features of features of features of features….

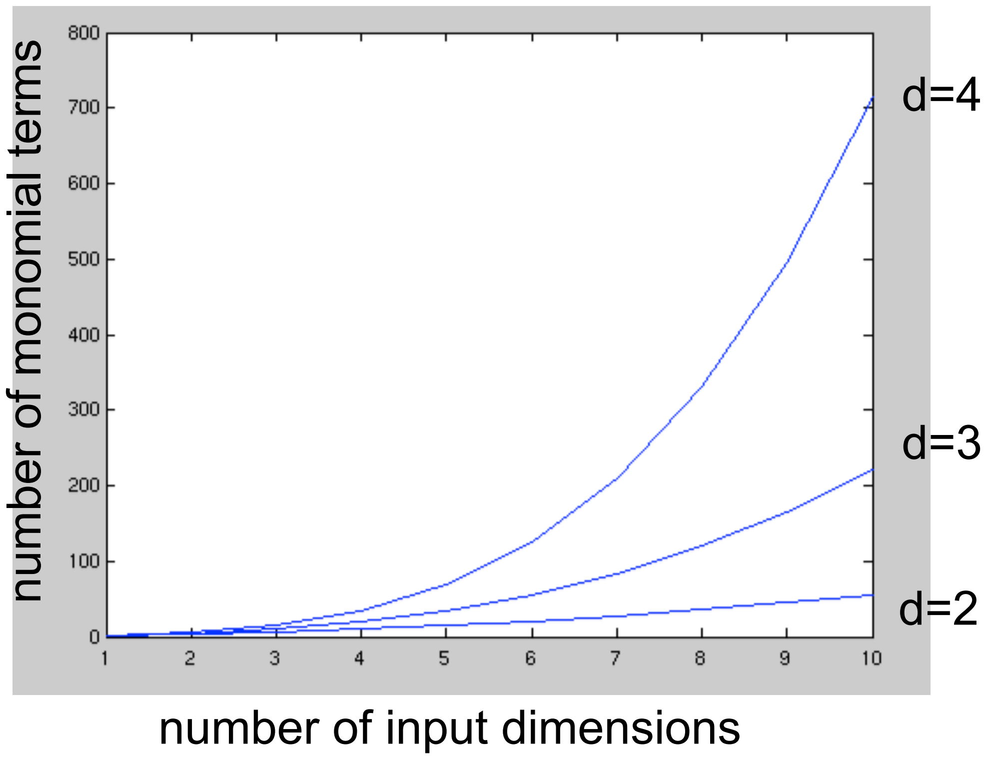


$$\phi(x) = \begin{pmatrix} x^{(1)} \\ \dots \\ x^{(n)} \\ x^{(1)}x^{(2)} \\ x^{(1)}x^{(3)} \\ \dots \\ e^{x^{(1)}} \\ \dots \end{pmatrix}$$

Feature space can get really large really quickly!

Higher order polynomials

$$\text{num. terms} = \binom{d+m-1}{d} = \frac{(d+m-1)!}{d!(m-1)!}$$



m – input features
d – degree of polynomial

grows fast!
d = 6, m = 100
about 1.6 billion terms

Dual formulation only depends on dot-products of the features!

$$\max_{\vec{\alpha}\geq 0,\ \sum_j \alpha_j y_j = 0} \sum_j \alpha_j - \frac{1}{2}\sum_{i,j} y_i y_j \alpha_i \alpha_j \left(\vec{x}_i \cdot \vec{x}_j\right)$$

First, we introduce a *feature mapping*:

$$\mathbf{x}_i \mathbf{x}_j \quad \rightarrow \quad \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$$

Next, replace the dot product with an equivalent *kernel* function:

$$\text{maximize}_\alpha \quad \sum_i \alpha_i - \tfrac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$$

$$K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$$

$$\sum_i \alpha_i y_i = 0$$

Polynomial kernel

d=1

$$\phi(u).\phi(v) = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} . \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = u_1v_1 + u_2v_2 = u.v$$

d=2

$$\phi(u).\phi(v) = \begin{pmatrix} u_1^2 \\ u_1u_2 \\ u_2u_1 \\ u_2^2 \end{pmatrix} . \begin{pmatrix} v_1^2 \\ v_1v_2 \\ v_2v_1 \\ v_2^2 \end{pmatrix} = u_1^2v_1^2 + 2u_1v_1u_2v_2 + u_2^2v_2^2$$

$$= (u_1v_1 + u_2v_2)^2$$

$$= (u.v)^2$$

For any *d (we will skip proof):*

$$\phi(u).\phi(v) = (u.v)^d$$

Polynomials of degree **exactly** *d*

Common kernels

- Polynomials of degree exactly *d*

$$K(\mathbf{u},\mathbf{v}) = (\mathbf{u}\cdot\mathbf{v})^d$$

- Polynomials of degree up to *d*

$$K(\mathbf{u},\mathbf{v}) = (\mathbf{u}\cdot\mathbf{v} + 1)^d$$

- Gaussian kernels

$$K(\vec{u},\vec{v}) = \exp\left(-\frac{||\vec{u}-\vec{v}||_2^2}{2\sigma^2}\right)$$

- Sigmoid

$$K(\mathbf{u},\mathbf{v}) = \tanh(\eta\mathbf{u}\cdot\mathbf{v} + \nu)$$

- And many others: very active area of research!